Webology (ISSN: 1735-188X)
Volume 18, Number 5, 2021

Intelligent Based Data Mining System For Medical Database Using Multi-Agent System

S. Sulaiha Beevi[1] , Dr. K.L. Shunmuganathan[2] , Dr. K. Manivannan[3]

[1]Research Scholar, Bharathiyar University Coimbatore.

[2]Deputy Director, Aarupadai Veedu Institute of Technology Vinayaka Mission's Research Foundation, Salem, Tamilnadu, India Chennai.

[3]Director Indstry Academia relations, Aarupadai Veedu Institute of Technology, Vinayaka Mission's Research Foundation, Salem, Tamilnadu, India.

ABSTRACT

Now-a-days, lifestyle disorder leads to many diseases in human being such as pressure, diabetes cancer and heart attack. The accurate prediction of these diseases through symptoms becomes difficult for physicians. The prediction of diseases at early stage is a challenging task. To overcome the above problem, data mining plays an important role for predicting the disease at early stage. Machine learning (ML) algorithms discover hidden information from disease dataset. In this thesis, the major affecting disease such as lung cancer is taken for the study and predict lung cancer at early stage through multi-agent-based architecture based on machine-learning. For predicting lung cancer at early stage, three different methods are proposed in this thesis. Initially, lung cancer is predicted through Logistic Regression (LR), k Nearest Neighbor (kNN), Naive Bayes (NB) and Support Vector Machine (SVM) methods. In this thesis, lung cancer prediction at earlier stage is performed with above three methods such asnon-Agent based lung dataset is applied with ML algorithms, Agent based lung dataset is applied with ML algorithms and Agent based lung dataset is applied with hybrid ML algorithms. In second method, machine learning is used in multi-agent platform for predicting lung cancer at early stage. Java Agent Development Environment based Multi Agent System (JADE-MAS) is developed for the prediction of lung cancer. In JADE-MAS, agent class is a superclass which allows the users to create agents. The JADE-MAS consists of three agents such as the medical practitioner Agent, & Classifier Agent and the Database Agent. Database Agent stores the dataset related to the patient. In Medical practitioner Agent, symptoms / risk factors are given as input to the dataset physician or patient. Finally, the agent-based lung dataset is processed with Classifier agent. In third method, JADE-MAS is applied with novel hybrid classifiers such as Multiple Linear Regression+ k Nearest Neighbour (MLR+kNN), Gaussian Kernel Support Vector Machine + Linear Regression (GKSVM +LR) and Gaussian Kernel Support Vector Machine + k Nearest Neighbour (GKSVM+kNN) for lunger cancer prediction at early stage.In this thesis, agent based and non-agent-based classification for lung cancer prediction at early stage is analyzed with sensitivity, specificity and accuracy. From the analysis, the optimum method for detecting lung cancer at early stage is identified and their advantages and disadvantages are studied.

Keywords: lung cancer, machine learning, Java Agent Development Environment based Multi Agent System (JADE-MAS), hybrid methods

1. Introduction

Cancer cells are cells that divide relentlessly, forming solid tumors or flooding the blood with abnormal cells. Normally human cells grow and divide to form new cells as required by body. Cells grow old or become damaged, they die, and new cells are formed. Cancer cells divide without stopping, grows and called as tumor. A cancer cell has thousands of mutations, certain number of these genetic changes in cancer cells and causes cancer cells to divide and grow. Mutations is the growth of cancer cells and referred as "driver mutations," whereas other mutations are considered "passenger mutations." Normal genes called proto-oncogenes and becomes "oncogenes", The mutation and code for proteins drive the growth of cancer and cancer lead to mortality. Tumor suppressor genes slows down the growing, repair damaged DNA, or makes cells to die and stops growing of cancer cells. The top 10 most common cancers are Breast cancer, Lung cancer, Prostate cancer, Colon cancer, Bladder cancer, Melanoma, Non-Hodgkin lymphoma, Thyroid Cancer, Kidney cancer and Leukemia. Lung's cancer occurs in the people who smoke. Lungs are the primary organs of respiratory system in humans which function to take in oxygen and expel carbon dioxide. During breathing air enters through mouth or nose and goes into the lungs through the trachea (windpipe). The trachea divides into two branches called as bronchi and enters into the lungs and further divides into smaller bronchi and smaller branches called bronchioles. At the end of each bronchioles tiny air sacs are present and known as alveoli. The alveoli absorb oxygen into the blood during inhale of air and remove carbon dioxide from the blood during exhale. Lung cancers start in the cells lining the bronchi and parts of the lung such as the bronchioles or alveoli. The function of lungs in the respiratory system is to extract oxygen from inhaled air and transfer to blood, and release carbon dioxide from the blood and this process is called as gas exchange. The inhaled air is a mixture of oxygen and other gases. In the throat, the trachea, or windpipe, filters the air. The trachea branches into two bronchi, tubes that lead to the lungs takes the air for filtering process. Two major types of lung cancer are non-small cell lung cancer and small cell lung cancer. Signs and symptoms of lung cancer are continuous coughing, coughing with blood, (small amount), Shortness of breath, Chest pain, Hoarseness, losing weight without trying, Bone pain and Headache. Several factors increase lung cancer. Risk factors can be controlled. Risk factors for lung cancer include Smoking, Passive Smoking, Previous radiation therapy, Exposure to radon gas, Exposure to asbestos and other carcinogens and Family history of lung cancer. In this paper, the first phase involves the analysis of different machine learning techniques such as Logistic Regression (LR), k Nearest Neighbour (kNN), Naive Bayes (NB), Support Vector Machine (SVM). The input lung cancer dataset is processed by this agentless method. The second phase includes the development of a Multi-Agent System (MAS) which analyses the different machine learning techniques in the similar dataset. This phase resulted in the implementation of a prototype with the best lung cancer detection technique. The third phase shows or implement the proposed hybrid machine learning technique on different dataset. The last phase uses the proposed architecture to implement genomic data analysis.

Problem statement

The lung dataset-based datamining provides the information about cancer and help physician for discovering hidden information. Till now the discovering of information from database performed through various algorithms such as SVM, k-NN, LR and NB. However, researchers need to focus on agent-based data collection of lung diseases. The lung cancer disease symptoms vary from region to region and due to change in food habits and lifestyle modification according to demography. Similarly, each machine learning algorithms have their own advantage and disadvantage. Researchers need to focus on hybrid algorithms.

Objectives

The principal objective of the research work is the design of concept to implement intelligent data mining system for lung databases through multi-agent system. The objective of this thesis is:

- To study, lung cancer disease pattern based on database collection such as Standard dataset and Agent based data collection.
- To analyses, lung cancer dataset based on multi-agent system for earlier lung cancer detection.
- To propose, Hybrid algorithms for lung dataset such as (i) LR + kNN (ii) GKSVM + LR (iii) GKSVM + kNN
- To analyse, Hybrid algorithms in Agent based data collection and lung dataset for discovering hidden information through various measures such as Precision, Recall and True Negative.
- To study, advantages and disadvantages of proposed method in standard dataset and agent-based lung cancer dataset.

2. Literature survey

Data Mining is applied in various fields such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile computing [1]. Data mining is applied in healthcare for detecting fraud and abuse cases [2]. Clinical decisions are often made based on physician intuition and experience rather than knowledge or information from collected dataset. The above practice leads to improper treatment, diagnosis, adverse drug effect, huge medical cost, and less quality of service to patients [3]. Data modeling and analysis help physician improving the quality of clinical decisions. Healthcare organizations perform data mining for meeting their long-term variant of disease and their symptoms [4].Lung's cancer occurs in the people who smoke. Lungs are the primary organs of respiratory system in humans which function to take in oxygen and expel carbon dioxide. Lungs are situated within the thoracic cavity of the chest. Always right lung is bigger than the left, and located below the heart [5]. Non-small cell lung cancer: non-small cell lung cancer is an umbrella term for several types of lung cancers such as squamous cell carcinoma, adenocarcinoma, and large cell carcinoma [6]. Small cell lung cancer: small cell lung cancer occurs in the people who have chain smoking habit [7]. Data Mining (DM) brings out hidden, valid, and potentially useful information and patterns in huge lung datasets [8-9]. DM process is divided into two parts i.e., Data Pre-processing and Data Mining. Data Pre-processing involves data cleaning, data integration, data reduction, and data transformation [10]. Healthcare data analysis is currently a challenging and crucial research for the development of a robust disease diagnosis at early-stage prediction system and prognosis. Classification algorithms improve disease detection automatically through supervised and unsupervised learning. Numerous machine learning algorithms have been developed and extract useful patterns from medical data over the years [11]. The patterns have been identified or disease prediction using classification and clustering algorithms. Researchers focuses on employing data mining in medical dataset for prediction of a broad range of diseases, including breast cancer [12], heart diseases, Parkinson's disease [13], hepatitis, and diabetes, only to name a few.Logistic Regression(LR) uses a logistic function model with binary dependent variable[14].k-NN is a type of instance-based learning, where function is only approximated locally and all computation is deferred until function evaluation [15].

Interference from literature survey

The effectiveness of classification and recognition systems has improved in a great deal to help medical experts in multiple diagnosing diseases. This literature survey reviewed classification methods, hybrid classification methods and agent-based classification methods for recognition of multiple diseases. The proposed system will focus on the multiple disease prediction with above mentioned three categories. Among these categories agent-based system with classifier provides better results than the other methods.

3. Methodology

Figure 1 shows the block diagram of hybrid methods with machine learning for lung cancer dataset. Lung cancer predicted through different algorithms such as machine learning (logistic regression, naïve bayes, KNN and support vector machine), Java Agent Development Environment Based Multi Agent System (JADE-MAS) and hybrid algorithms (Multiple Linear Regression+ k Nearest Neighbour (MLR+kNN), Gaussian Kernel Support Vector Machine + Linear Regression (GKSVM +LR) and Gaussian Kernel Support Vector Machine + k Nearest Neighbour (GKSVM+kNN)). The three methods are proposed methods for lung cancer prediction. Hybrid algorithms gives more accuracy compared than other methods to predict lung cancer.

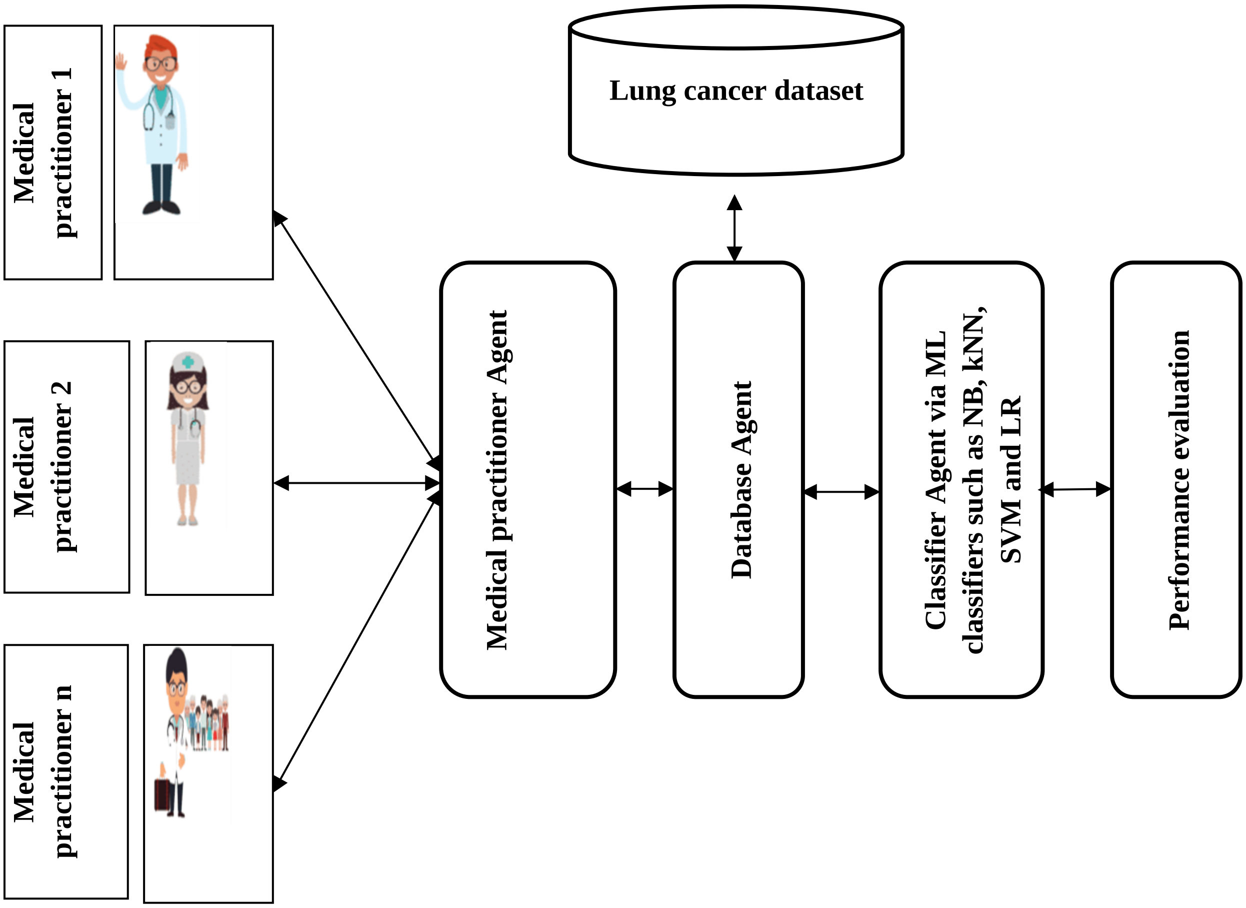


Fig.1 shows block diagram of hybrid algorithms for lung cancer

3.1.Machine learning

Machine Learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. ML algorithms build a model based on sample data, known as "training data", to make predictions or decisions without being explicitly programmed to do so. Classification is one of the most important aspects of supervised learning. Various classification algorithms in machine learning such as Logistic Regression (LR), k Nearest Neighbour (kNN), Naive Bayes (NB), and Support Vector Machine (SVM).

3.1.1. Logistic regression

Logistic Regression (LR) is basically a supervised classification algorithm. In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”. assumes that the data follows a linear function, Logistic regression models the data using the sigmoid function.

3.1.2. K-Nearest Neighbor

k-nearest Neighbors (kNN) Algorithm is a non-parametric method. In k-NN classification, the output is a class membership. A data is classified by a plurality vote of its neighbors, with the data being assigned to the class most common among its k nearest neighbors (k is a positive integer , typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance. kNN runs this formula to compute the distance between each data point and the test data. It then finds the probability of these points being similar to the test data and classifies it based on which points share the highest probabilities. The best choice of k depends upon the data; generally, a larger value of k reduces effect of the noise on the classification but make boundaries between classes less distinct. The special case where the class is predicted to be the class of the closest training sample (i.e., when k = 1) is called the NN algorithm. Popular way of choosing the empirically optimal k in this setting is via bootstrap method. With the help of KNN is can easily identify the category or class of a particular dataset.

3.1.3. Naïve bayes

A Naive Bayes (NB) classifier is a probabilistic machine learning model that is used for classification task. Assume that the value of a particular feature is independent of the value of any other feature, given the class variable. NB is used to compute posterior probabilities given observations. For example, a patient may be observed to have certain symptoms. Bayes theorem can be used to compute the probability that a proposed diagnosis is correct, given that observation. In simple terms, a NB classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature.

3.1.4. Support Vector Machine

Support Vector Machine (SVM) model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. SVM takes the form of mapping input space into higher dimensional space to support nonlinear classification problems where the maximum separation of the hyperplane is constructed. The hyperplane is a linear pattern (depends on the kernel) that maximizes the margin resulting maximum value between classification classes.

3.2.Java Agent Development Environment Based Multi Agent System (JADE-MAS)

JADE-MAS is a middleware which facilitates the development of multi-agent systems under the standard Foundation for Intelligent Physical Agents (FIPA) for which purpose it creates multiple containers for agents, each of them can run on one or more systems. JADE-MAS are a distributed agent's platform, which has a container for each host where you are running the agents. Each platform must have a parent container that has two special agents called AMS and Directory Facilitator (DF).Directory Facilitator (DF) gives a directory which announces which agents such as medical practitioner agent, database agent, and Database Agent are available on the platform. To access the DF agent the class "jade. domain. DF Service" and its static methods such as register, deregister, modify and Search are used correctly. Agent Management System (AMS) controls the platform. It is the only one who can create and destroy other agents, destroy containers, and stop the platform. AMS Service an agent is created which automatically runs the register method of the AMS by default before executing the method setup from the new agent. When an agent is destroyed it executes its takedown () method by default and automatically calls the deregister method of the AMS.

3.3.Hybrid classifiers

Hybrid classification system is proposed to diagnose lung cancer by combining individual classifiers. Initially, Java Agent Development Environment based Multi Agent System (JADE-MAS) for the modelling and simulation of lung cancer dataset. From this JADE-MAS model, classifiers such as Multiple Linear Regression+ k Nearest Neighbour (MLR+kNN), Gaussian Kernel Support Vector Machine + Linear Regression (GKSVM +LR) and Gaussian Kernel Support Vector Machine + k Nearest Neighbour (GKSVM+kNN) have been implemented for lung cancer diagnosis.

3.3.1. Multiple Linear Regression+ k Nearest Neighbour (MLR+kNN)

MLR+kNN strategy is proposed to combine both Multiple Linear Regression (MLR) and k-Nearest Neighbor (kNN) algorithm in an efficient way for lung cancer classification. The hybrid MLR+kNN classifier perform the classification of cancer data to reduce the error of the classification, and multiple regressions are used to extract label-dependent information from the label space. The multi-label classifier incorporates the label dependency in the label space and space for prediction. The objective of this MLR+kNN classifier is to classify a data into more than one class instead of a single class. kNN and MLR is combined for classification label set of a data. Multiple Linear Regression (MLR) is an extension of simple Linear Regression for incorporating the dependencies of more than one variable

3.3.2. Gaussian Kernel Support Vector Machine + Linear Regression (GKSVM+LR)

A binary classifier dataset (D) is a mapping, D: $\mathrm{A}^{\mathrm{m}} \rightarrow [0,1]$, where $\mathrm{S}^{\mathrm{m}}$is the m-dimensional space containing real number, binary as well as category. For a vector x ∈ A^{m},D(x) can be considered as a probability function, whose value is the probability that x is labelled as positive. Suppose that S and L are the mappings made up by GKSVM and LR respectively, take a new mapping, the composition of S and L, as D. GKSVM classifier uses a technique called the kernel trick in which kernel takes a low dimensional input space and transforms it into a higher dimensional space. In simple words, kernel converts non-separable problems into separable problems by adding more dimensions to it. It makes GKSVM more powerful, flexible, and accurate.

3.3.3. Gaussian Kernel Support Vector Machine + k Nearest Neighbour (GKSVM+kNN)

Hybrid classifier is performed by incorporating the Gaussian Kernel Support Vector Machine (GKSVM) learning information into the k-Nearest Neighbor (kNN) classifier. The GKSVM is well known for its extraordinary generalization capability even with limited lung cancer dataset samples, and it is very useful for remote sensing applications as lung cancer dataset samples are usually limited. The kNN has been widely used in lung cancer classification due to its simplicity and effectiveness. However, the kNN is instance-based and needs to keep all the training samples for classification, which could cause not only high computation complexity but also overfitting issues. Meanwhile, the performance of the kNN classifier is sensitive to the neighborhood size k and how to select the value of the parameter k relies heavily on practice and experience. Based on the observations that the GKSVM can contribute to the kNN on the problems of smaller training samples size as well as the selection of the parameter k, proposed GKSVM with k-Nearest Neighbor (abbreviated as GKSVM-kNN) hybrid classification approach which can simplify the parameter selection while maintaining classification accuracy. The proposed approach is consisting of two stages. In the first stage, the GKSVM is performed on the training samples to obtain the reduced Support Vectors (SVs) for each of the sample categories. In the second stage, a k-Nearest Neighbor (kNN) classifier is used to classify a testing sample, i.e., the average Euclidean distance between the testing data point to each set of vectors from different categories is calculated and the kNN identifies the category with minimum distance.

4. Results and discussion

The results are measured with respect to three datasets such as Lung Cancer, Hungarian and Covid 19 Symptoms dataset. Hungarian dataset consists of 294 samples with 76 attributes. Among them 14 attributes only used. Attribute Information, 1. #3 (age) , 2. #4 (sex), 3. #9 (cp) , 4. #10 (treetops),5. #12 (Chol), 6. #16 (fbs), 7. #19 (restecg), 8. #32 (thalach), 9. #38 (exang), 10. #40 (oldpeak), 11. #41 (slope), 12. #44 (ca), 13. #51 (thal) and 14. #58 (num) (the predicted attribute) . This Hungarian dataset is collected from the repository of University of California at Irvine (UCI). Covid 19 Symptoms dataset is collected from https://www.kaggle.com/iamhungundji/covid19-symptoms-checker. From those we have considered 285 samples with 27 attributes, 27 attributes consist of 23 belongs to normal attributes and 4 belongs to class. Those attributes are Fever, Tiredness, Dry-Cough, Difficulty-in-Breathing, Sore-Throat, None_Symptoms, Pains, Nasal-Congestion, Runny-Nose, Diarrhea, None_ Experiencing, Age_0-9, Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender_Female, Gender_Male, Gender_Transgender,Severity_Mild, Severity_Moderate, Severity_Severe, Severity_None, Contact_Dont-Know, Contact_No, Contact_Yes, and Country.

The results are measured with respect to precision, recall, F-measure, and accuracy are applied to the diagnosis condition can get sufficient explanations (See Table 1-3).

Table 1 Performance Comparison Analysis of Lung Cancer Dataset

Lung cancer dataset- Results (%)				
Classifiers	**Precision**	**Recall**	**F-measure**	**Accuracy**
LR	71.57	71.85	71.19	71.88
kNN	74.13	74.52	73.94	75.00
SVM	77.47	77.47	77.33	78.12
JADE-MAS +LR	71.20	71.11	72.51	71.88
JADE-MAS+ kNN	74.90	74.44	74.56	75.00
JADE-MAS + SVM	77.47	77.47	77.33	78.13
JADE-MAS- MLR+kNN	87.07	87.07	87.07	87.50
JADE-MAS- GKSVM +LR	94.10	93.93	94.02	93.75
JADE-MAS- GKSVM+kNN	97.43	96.67	97.05	96.88

Table 2.Performance Comparison Analysis of Hungarian Heart Disease Dataset

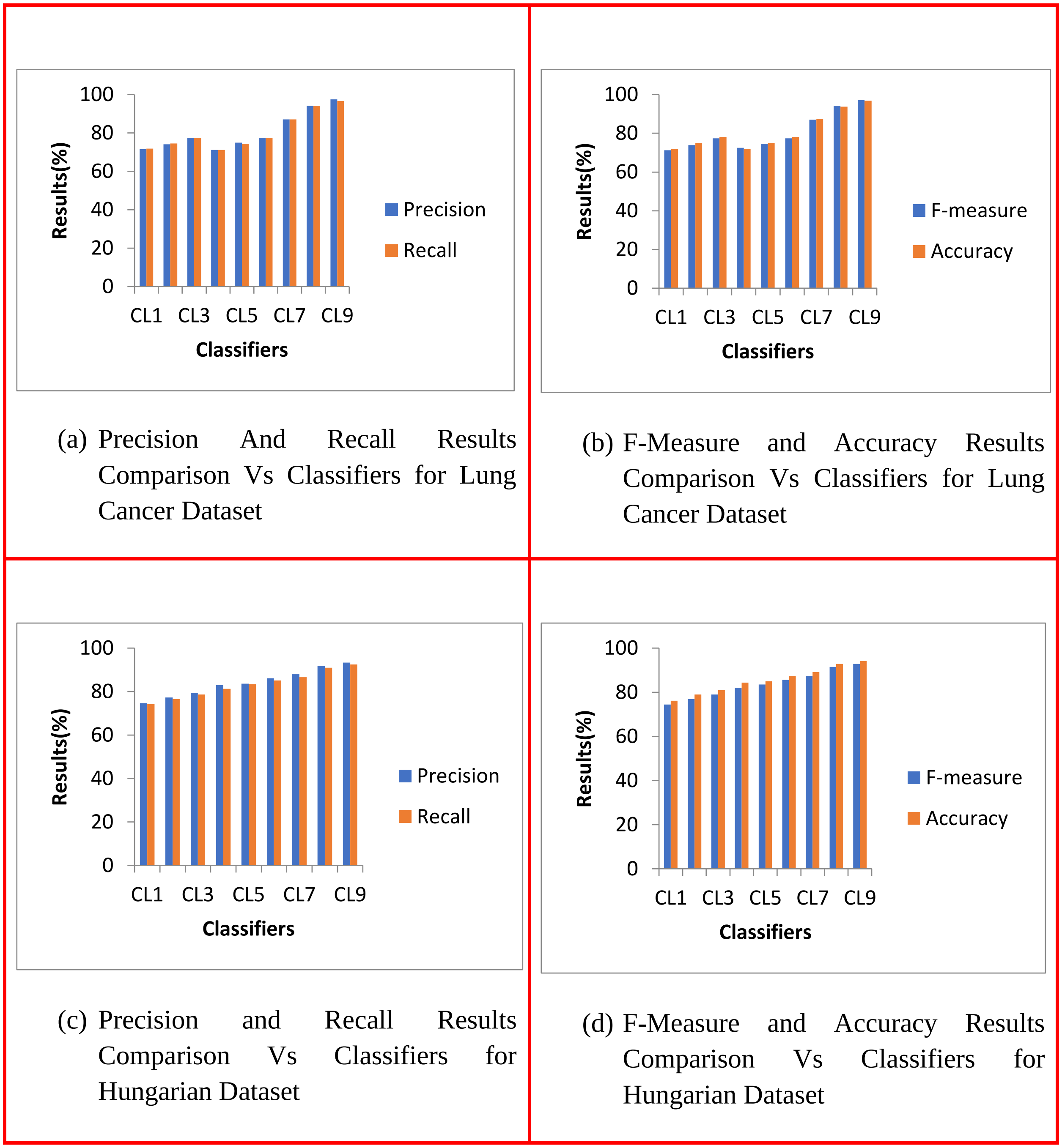
Hungarian dataset- Results (%)				
Classifiers	**Precision**	**Recall**	**F-measure**	**Accuracy**

LR	74.70	74.23	74.47	76.19
kNN	77.27	76.57	76.92	78.91
SVM	79.34	78.58	78.96	80.95
JADE-MAS +LR	82.97	81.24	82.09	84.35
JADE-MAS+ kNN	83.63	83.42	83.52	85.03
JADE-MAS + SVM	86.12	85.07	85.59	87.41
JADE-MAS- MLR+kNN	88.01	86.61	87.31	89.12
JADE-MAS- GKSVM +LR	91.87	90.97	91.42	92.86
JADE-MAS- GKSVM+kNN	93.30	92.45	92.88	94.22

Table 7.3. Performance Comparison Analysis of Covid 19 Symptoms Dataset

Covid 19 symptoms dataset- Results (%)				
Classifiers	**Precision**	**Recall**	**F-measure**	**Accuracy**
LR	76.19	76.88	76.08	76.14
kNN	78.59	78.81	78.62	78.59
SVM	80.37	80.57	80.39	80.35
JADE-MAS +LR	83.13	83.77	83.32	83.16
JADE-MAS+ kNN	84.61	85.27	84.68	84.56
JADE-MAS + SVM	87.04	87.37	87.00	87.02
JADE-MAS- MLR+kNN	89.48	89.62	89.51	89.47
JADE-MAS- GKSVM +LR	91.86	91.94	91.86	91.93
JADE-MAS- GKSVM+kNN	94.39	94.47	94.42	94.38

Figure 2 shows the output of hybrid algorithms for lung cancer.(a) Precision And Recall Results Comparison Vs Classifiers for Lung Cancer Dataset. (b) F-Measure and Accuracy Results Comparison Vs Classifiers for Lung Cancer Dataset. (c) Precision and Recall Results Comparison Vs Classifiers for Hungarian Dataset. (d) F-Measure and Accuracy Results Comparison Vs Classifiers for Hungarian Dataset. (e) Precision And Recall Results Comparison Vs Classifiers for COVID 19 Dataset. (f)F-Measure and Accuracy Results Comparison Vs. Classifiers for COVID 19 Dataset.

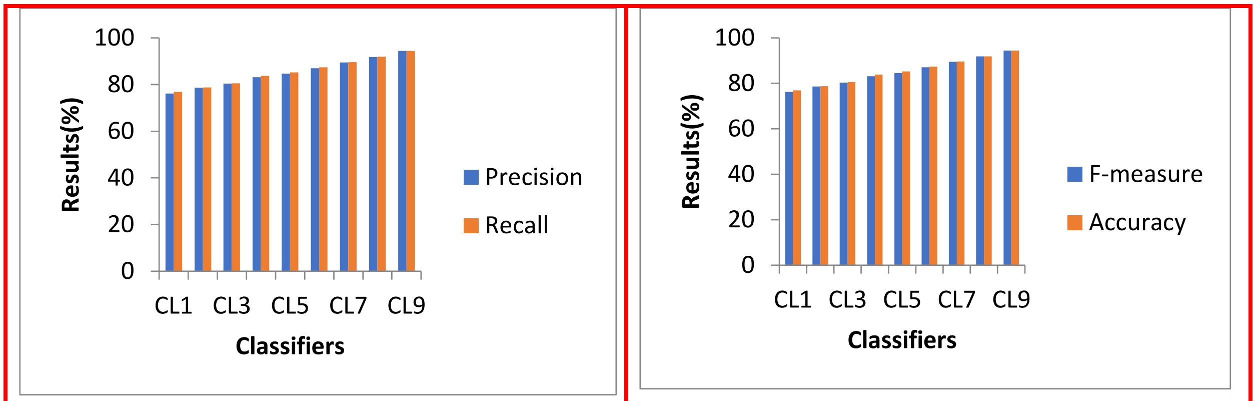


(a) Precision And Recall Results Comparison Vs Classifiers for Lung Cancer Dataset

(b) F-Measure and Accuracy Results Comparison Vs Classifiers for Lung Cancer Dataset

(c) Precision and Recall Results Comparison Vs Classifiers for Hungarian Dataset

(d) F-Measure and Accuracy Results Comparison Vs Classifiers for Hungarian Dataset



(e) Precision And Recall Results Comparison Vs Classifiers for COVID 19 Dataset

(f) F-Measure and Accuracy Results Comparison Vs. Classifiers for COVID 19 Dataset

Fig 2 shows output of three dataset with hybrid methods

Figure 2 (a) shows the precision and recall results comparison of various classifiers such as CL1-LR, CL2-kNN, CL3-SVM, CL4-JADE-MAS+LR, CL5-JADE-MAS+ kNN, CL6-JADE-MAS+SVM, CL7-JADE-MAS-MLR+kNN, CL8-JADE-MAS- GKSVM +LR, and CL9-JADE-MAS-GKSVM+kNN. From the results it concludes that the final GKSVM+kNN classifier gives higher precision results of 97.43%, whereas other classifiers from one to eight gives precision results of 71.57%,74.13%, 77.47%, 71.20%, 74.90%, 77.47%, 87.07% and 94.10% respectively. It increases the positive results for prediction. From the results it concludes that the final GKSVM+kNN classifier gives higher recall results of 96.67%, whereas other classifiers from one to eight gives precision results of 71.85%,74.52%, 77.47%, 71.11%, 74.44%, 77.47%, 87.07% and 93.93% respectively (see Table 1).

Figure 2 (b) shows the accuracy and F-measure results of various classifiers with respect to lung cancer. From the results it concludes that the final GKSVM+kNN classifier gives higher F-measure results of 97.05%, whereas other classifiers from one to eight gives F-measure results of 71.19%, 73.94%, 77.33%, 72.51%, 74.56%,77.33%, 87.07% and 94.02% respectively. From the results it concludes that the final GKSVM+kNN classifier gives higher accuracy results of 96.88%, whereas other classifiers from one to eight gives accuracy results of 71.88%, 75.00%, 78.12%, 71.88%, 75.00%, 78.13%, 87.50% and 93.75% respectively (see Table 1).

Figure 2 (c) shows the precision and recall results comparison of various classifiers with respect to Hungarian dataset. From the results it concludes that the last GKSVM+kNN classifier gives higher precision results of 93.30%, whereas other classifiers from one to eight gives precision results of 74.70%,77.27%, 79.34%, 82.97%, 83.63%, 86.12%, 88.01% and 91.87% respectively. From the results it concludes that the final GKSVM+kNN classifier gives higher recall results of

92.45%, whereas other classifiers from one to eight gives precision results of 74.23%,76.57%, 78.58%, 81.24%, 83.42%, 85.07%, 86.61% and 90.97% respectively (See Table 2).

Figure 2 (d) shows the accuracy and F-measure results of various classifiers with respect to Hungarian cancer. From the results it concludes that the final GKSVM+kNN classifier gives higher F-measure results of 92.88%,whereas other classifiers from one to eight gives F-measure results of 74.47%, 76.92%, 78.96%, 82.09%, 83.52%,85.59%, 87.31% and 91.42% respectively. From the results it concludes that the final GKSVM+kNN classifier gives higher accuracy results of 94.22%, whereas other classifiers from one to eight gives accuracy results of 76.19%, 78.91%, 80.95%, 84.35%, 85.03%, 87.41%, 89.12% and 92.86% respectively (See Table 2).

Figure 2 (e) shows the precision and recall results comparison of various classifiers with respect to Covid 19 dataset. From the results it concludes that the last GKSVM+kNN classifier gives higher precision results of 94.39%,whereas other classifiers from one to eight gives precision results of 76.19%, 78.59%, 80.37%, 83.13%, 84.61%, 87.04%, 89.48%, and 91.86% respectively. From the results it concludes that the final GKSVM+kNN classifier gives higher recall results of 94.47%, whereas other classifiers from one to eight gives precision results of 76.88%,78.81%, 80.57%, 83.77%, 85.27%, 87.37%, 89.62% and 91.94% respectively (See Table 3).

Figure 2 (f) shows the accuracy and F-measure results of various classifiers with respect to Covid 19 dataset. From the results it concludes that the final GKSVM+kNN classifier gives higher F-measure results of 94.42%, %, whereas other classifiers from one to eight gives F-measure results of 76.08%, 78.62%, 80.39%, 83.32%, 84.68%,87.00%, 89.51% and 91.86% respectively. From the results it concludes that the final GKSVM+kNN classifier gives higher accuracy results of 94.38%, whereas other classifiers from one to eight gives accuracy results of 76.14%, 78.59%, 80.35%, 83.16%, 84.56%, 87.02%, 89.47% and 91.93% respectively (See Table 3).

5. Conclusion and future work

Accurate and on time diagnosis of disease is important for heart failure prevention and treatment. The diagnosis of disease through traditional medical history has been considered as not reliable in many aspects. Multiple disease diagnosis through the machine-learning-based system has been reported in various research studies. This work will focus on the developing a several machine learning methods for multiple disease diagnosis in medical data mining. The initial contribution of the work will focus on the classifiers such as Logistic Regression (LR), k Nearest Neighbour (kNN), Naive Bayes (NB), and Support Vector Machine (SVM). Here the missing data is replaced by using min max normalization. It is one of the most common ways to normalize data. For every attribute, the minimum value of that attribute gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1. LR, target variable can take only discrete values for given set of features. kNN algorithm, data is classified by a plurality vote of its neighbors, with the data being assigned to the class most common among its k nearest neighbors. NB classifier is used to compute posterior probabilities given observations. Finally, these classifiers have been applied to multiple disease dataset. The second contribution of the work deals with the problem of multi-disease classification with agent framework. Java Agent Development Environment based Multi Agent System (JADE-MAS) is a middleware which facilitates the development of multi-agent systems for agents, each of them can run on one or more systems. This system is performed based on three agents such as the medical practitioner Agent, Classifier Agent and the Database Agent. Medical practitioner Agent enables the user to input his or her symptoms/risk factors in order for the classifier agent to classify a cancer. The classify agent (i.e., either LR, kNN, NB and SVM) is responsible for classifying the symptoms presented by the medical practitioner agent into either cancer or non-cancer using the classifiers. Database agent will maintain the data related to the patient. The final contribution of the work, JADE-MAS approach with hybrid classifiers such as Multiple Linear Regression+ k Nearest Neighbour (MLR+kNN), Gaussian Kernel Support Vector Machine + Linear Regression (GKSVM +LR) and Gaussian Kernel Support Vector Machine + k Nearest Neighbour (GKSVM+kNN) has been introduced for multiple disease classification. The hybrid MLR+kNN classifier perform the classification of data in order to reduce the error of the classification, and multiple regressions are used to extract label-dependent information from the label space. In MLR+kNN classifier, linear model is generated for each label by using the labels from the training set. Here kNN is used for finding the most similar k neighbors from training set. GKSVM+LR classifier, input dataset mappings are made up by GKSVM and LR respectively. In GKSVM +LR classifier, each GKSVM classifier related to a split dataset can perform classification independently. GKSVM+kNN classifier is performed by incorporating the GKSVM learning information into the kNN classifier. kNN has the problem of smaller training samples size as well as the selection of the parameter k, which is solved by using the GKSVM algorithm. Finally, these classifiers have been measured using the metrics like precision, recall, F-measure and Accuracy. Designing a decision support system through machine-learning-based method will be more suitable for diagnosis of multiple diseases. Additionally, some irrelevant features reduced the performance of the diagnosis system and increased the computation time. So, another innovative dimension of this study was the usage of feature selection algorithms to choose best features that improve the classification accuracy as well as reduce the execution time of the diagnosis system. In the future work, focus on more experiments in order to increase the performance of these predictive classifiers in disease diagnosis by using Feature Selection algorithms and optimization techniques.

REFERENCES

1. Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F. and Hua, L., 2012. Data mining in healthcare and biomedicine: a survey of the literature. Journal of medical systems, 36(4), pp.2431-2448.

2. Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.

3. Chang, C.L. and Chen, C.H., 2009. Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Systems with Applications, 36(2), pp.4035-4041.

4. Chauhan D. and V. Jaiswal, "An efficient data mining classification approach for detecting lung cancer disease," in Communication and Electronics Systems (ICCES), International Conference on, 2016, pp. 1–8.

5. Collins, L.G., Haines, C., Perkel, R. and Enck, R.E., 2007. Lung cancer: diagnosis and management. American family physician, 75(1), pp.56-63.

6. Zappa, C. and Mousa, S.A., 2016. Non-small cell lung cancer: current treatment and future advances. Translational lung cancer research, 5(3), pp.288.

7. Zimmerman, S., Das, A., Wang, S., Julian, R., Gandhi, L. and Wolf, J., 2019. 2017–2018 scientific advances in thoracic oncology: small cell lung cancer. Journal of Thoracic Oncology, 14(5), pp.768-783.

8. Das, H., Naik, B. and Behera, H.S., 2018. Classification of diabetes mellitus disease (DMD): a data mining (DM) approach. In Progress in computing, analytics and networking (pp. 539-549). Springer, Singapore.

9. Sohail, M.N., Jiadong, R., Uba, M.M. and Irshad, M., 2019. A comprehensive looks at data mining techniques contributing to medical data growth: a survey of researcher reviews. In Recent Developments in Intelligent Computing, Communication and Devices (pp. 21-26). Springer, Singapore.

10. Zhang, C., Cao, L. and Romagnoli, A., 2018. On the feature engineering of building energy data mining. Sustainable cities and society, 39, pp.508-518.

11. Ramos-Pollán R., M. Á. Guevara-López, and E. Oliveira, "A software framework for building biomedical machine learning classifiers through grid computing resources," Journal of Medical Systems, vol. 36, no. 4, pp. 2245–2257, 2012.

12. Malik A. and J. Iqbal, "Extreme learning machine based approach for diagnosis and analysis of breast cancer," Journal of the Chinese Institute of Engineers, vol. 39, no. 1, pp. 74–78, 2016.

13. Hariharan M., K. Polat, and R. Sindhu, "A new hybrid intelligent system for accurate detection of Parkinson's disease," Computer Methods and Programs in Biomedicine, vol. 113, no. 3, pp. 904–913, 2014.

14. Tolles, J. and Meurer, W.J., 2016. Logistic regression: relating patient characteristics to outcomes. Jama, 316(5), pp.533-534.

15. Shouman, M., Turner, T. and Stocker, R., 2012. Applying k-nearest neighbour in diagnosing heart disease patients. International Journal of Information and Education Technology, 2(3), pp.220-223.